(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 September 2005 (29.09.2005)

PCT

(10) International Publication Number
WO 2005/089246 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/US2005/008298

(22) International Filing Date: 11 March 2005 (11.03.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/552,653 13 March 2004 (13.03.2004) US
60/581,257 18 June 2004 (18.06.2004) US

(71) Applicant *(for all designated States except US)*: **CLUSTER RESOURCES, INC.** [US/US]; 622 North 900 East, Spanish Fork, UT 84660 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **JACKSON, David, Brian** [US/US]; 10890 S. Conestoga Drive, Spanish Fork, UT 84660 (US).

(74) Agent: **ISAACSON, Thomas, M.**; Law Office of Thomas M. Isaacson, 850 Lindy Lane, Huntingtown, MD 20639 (US).

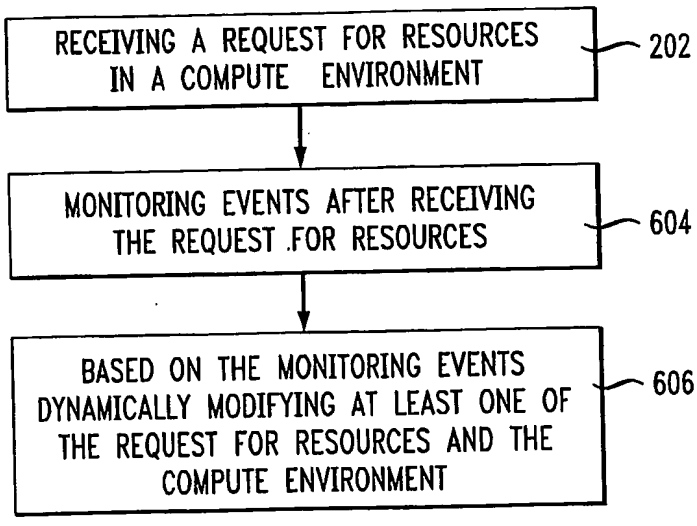
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR PROVIDIING ADVANCED RESERVATIONS IN A COMPUTE ENVIRONMENT



(57) Abstract: A system and method are disclosed for dynamically reserving resources within a cluster environment. The method embodiment of the invention comprises receiving a request for resources in the cluster environment, monitoring events after receiving the request for resources and based on the monitored events, dynamically modifying at least one of the request for resources and the cluster environment.

WO 2005/089246 A2

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHOD FOR PROVIDING ADVANCED RESERVATIONS IN A COMPUTE ENVIRONMENT

PRIORITY CLAIM

[0001] The present application claims priority to U.S. Provisional Application No. 60/552,653 filed March 13, 2004, the contents of which are incorporated herein by reference. The present application also cites priority to U.S. Provisional Application No. 60/581,257 filed June 18, 2004, the contents of which are incorporated herein by reference.

RELATED APPLICATIONS

[0002] The present application is related to Attorney Docket Numbers 010-0011A, 010-0011B, 010-0011C, 010-0013, 010-0019, 010-0026, 010-0028 and 010-0030 filed on the same day as the present application. The content of each of these cases is incorporated herein by reference.

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0003] The present invention relates to reservations in a compute environment and more specifically to a system and method of providing advanced reservations to resources within a compute environment such as a cluster.

2. Introduction

[0004] There are challenges in the complex process of managing the consumption of resources within a compute environment such as a grid, compute farm or cluster of computers. Grid computing may be defined as coordinated resource sharing and problem solving in dynamic, multi-institutional collaborations. Many computing projects require much more computational power and resources than a single computer may provide. Networked computers with peripheral resources such as printers, scanners, I/O devices, storage disks, scientific devices and instruments, etc. may need to be coordinated and utilized to complete a task. The term compute resource generally refers to computer processors, network bandwidth, and any of these peripheral resources as well. A compute farm may comprise a plurality of computers coordinated for such purposes of handling Internet traffic. The web search website Google® had a compute farm used to process its network traffic and Internet searches.

[0005] Grid/cluster resource management generally describes the process of identifying requirements, matching resources to applications, allocating those resources, and scheduling and monitoring grid resources over time in order to run grid applications or jobs submitted to the compute environment as efficiently as possible. Each project or job will utilize a different set of

resources and thus is typically unique. For example, a job may utilize computer processors and disk space, while another job may require a large amount of network bandwidth and a particular operating system. In addition to the challenge of allocating resources for a particular job or a request for resources, administrators also have difficulty obtaining a clear understanding of the resources available, the current status of the compute environment and available resources, and real-time competing needs of various users. One aspect of this process is the ability to reserve resources for a job. A cluster manager will seek to reserve a set of resources to enable the cluster to process a job at a promised quality of service.

[0006] General background information on clusters and grids may be found in several publications. See, e.g., Grid Resource Management, State of the Art and Future Trends, Jarek Nabrzyski, Jennifer M. Schopf, and Jan Weglarz, Kluwer Academic Publishers, 2004; and Beowulf Cluster Computing with Linux, edited by William Gropp, Ewing Lusk, and Thomas Sterling, Massachusetts Institute of Technology, 2003.

[0007] It is generally understood herein that the terms grid and cluster are interchangeable, although they have different connotations. For example, when a grid is referred to as receiving a request for resources and the request is processed in a particular way, the same method may also apply to other compute environments such as a cluster or a compute farm. A cluster is generally defined as a collection of compute nodes organized for accomplishing a task or a set of tasks. In general, a grid will comprise a plurality of clusters as will be shown in FIG. 1A. Several general challenges exist when attempting to maximize resources in a grid. First, there are typically multiple layers of grid and cluster schedulers. A grid 100 generally comprises a group of clusters or a group of networked computers. The definition of a grid is very flexible and may mean a number of different configurations of computers. The introduction here is meant to be general given the variety of configurations that are possible. A grid scheduler 102 communicates with a plurality of cluster schedulers 104A, 104B and 104C. Each of these cluster schedulers communicates with a respective resource manager 106A, 106B or 106C. Each resource manager communicates with a respective series of compute resources shown as nodes 108A, 108B, 108C in cluster 110, nodes 108D, 108E, 108F in cluster 112 and nodes 108G, 108H, 108I in cluster 114.

[0008] Local schedulers (which may refer to either the cluster schedulers 104 or the resource managers 106) are closer to the specific resources 108 and may not allow grid schedulers 102 direct access to the resources. The grid level scheduler 102 typically does not own or control the actual resources. Therefore, jobs are submitted from the high level grid-scheduler 102 to a local set of resources with no more permissions that then user would have. This reduces efficiencies and can render the reservation process more difficult.

[0009] The heterogeneous nature of the shared compute resources also causes a reduction in efficiency. Without dedicated access to a resource, the grid level scheduler 102 is challenged with the high degree of variance and unpredictability in the capacity of the resources available for use. Most resources are shared among users and projects and each project varies from the other. The performance goals for projects differ. Grid resources are used to improve performance of an application but the resource owners and users have different performance goals: from optimizing the performance for a single application to getting the best system throughput or minimizing response time. Local policies may also play a role in performance.

[0010] Within a given cluster, there is only a concept of resource management in space. An administrator can partition a cluster and identify a set of resources to be dedicated to a particular purpose and another set of resources can be dedicated to another purpose. In this regard, the resources are reserved in advance to process the job. There is currently no ability to identify a set of resources over a time frame for a purpose. By being constrained in space, the nodes 108A, 108B, 108C, if they need maintenance or for administrators to perform work or provisioning on the nodes, have to be taken out of the system, fragmented permanently or partitioned permanently for special purposes or policies. If the administrator wants to dedicate them to particular users, organizations or groups, the prior art method of resource management in space causes too much management overhead requiring a constant adjustment the configuration of the cluster environment and also losses in efficiency with the fragmentation associated with meeting particular policies.

[0011] To manage the jobs submissions or requests for resources within a cluster, a cluster scheduler will employ reservations to insure that jobs will have the resources necessary for processing. Figure 1B illustrates a cluster/node diagram for a cluster 124 with nodes 120. Time is along the X axis. An access control list 114 (ACL) to the cluster is static, meaning that the ACL is based on the credentials of the person, group, account, class or quality of service making the request or job submission to the cluster. The ACL 114 determines what jobs get assigned to the cluster 110 via a reservation 112 shown as spanning into two nodes of the cluster. Either the job can be allocated to the cluster or it can't and the decision is determined based on who submits the job at submission time. The deficiency with this approach is that there are situations in which organizations would like to make resources available but only in such a way as to balance or meet certain performance goals. Particularly, groups may want to establish a constant expansion factor and make that available to all users or they may want to make a certain subset of users that are key people in an organization and want to give them special services but only when their response time drops below a certain threshold. Given the prior art model, companies are unable to have the flexibility over their cluster resources.

[0012] To improve the management of compute resources, what is needed in the art is a method for a scheduler, such as a grid scheduler, a cluster scheduler or cluster workload management system to manage resources more efficiently. Furthermore, given the complexity of the cluster environment, what is needed is more power and flexibility in the reservations process.

SUMMARY OF THE INVENTION

[0013] Additional features and advantages of the invention will be set forth in the description which follows, and in part will be obvious from the description, or may be learned by practice of the invention. The features and advantages of the invention may be realized and obtained by means of the instruments and combinations particularly pointed out in the appended claims. These and other features of the present invention will become more fully apparent from the following description and appended claims, or may be learned by the practice of the invention as set forth herein.

[0014] The invention relates to systems, methods and computer-readable media for dynamically modifying either compute resources or a reservation for compute resources within a compute environment such as a grid or a cluster. In one aspect of the invention, a method of dynamically modifying resources within a compute environment comprises receiving a request for resources in the compute environment, monitoring events after receiving the request for resources and based on the monitored events, dynamically modifying at least one of the request for resources and the compute environment.

[0015] The invention enables an improved matching between a reservation and jobs submitted for processing in the compute environment. A benefit of the present invention is that the compute environment and the reservation or jobs submitted under the reservation will achieve a better fit. The closer the fit between jobs, reservations and the compute resources provides increased efficiency of the resources.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] In order to describe the manner in which the above-recited and other advantages and features of the invention can be obtained, a more particular description of the invention briefly described above will be rendered by reference to specific embodiments thereof which are illustrated in the appended drawings. Understanding that these drawings depict only typical embodiments of the invention and are not therefore to be considered to be limiting of its scope, the invention will be described and explained with additional specificity and detail through the use of the accompanying drawings in which:

[0017] FIG. 1A illustrates generally a grid scheduler, cluster scheduler, and resource managers interacting with compute nodes within plurality of clusters;

[0018] FIG. 1B illustrates an access control list which provides access to resources within a compute environment;

[0019] FIG. 2A illustrates a plurality of reservations made for compute resources;

[0020] FIG. 2B illustrates a plurality of reservations and jobs submitted within those reservations;

[0021] FIG. 3 illustrates a dynamic access control list;

[0022] FIG. 4 illustrates a reservation creation window;

[0023] FIG. 5 illustrates a dynamic reservation migration process;

[0024] FIG. 6 illustrates a method embodiment of the invention; and

[0025] FIG. 7 illustrates another method aspect of the invention.

DETAILED DESCRIPTION OF THE INVENTION

[0026] Various embodiments of the invention are discussed in detail below. While specific implementations are discussed, it should be understood that this is done for illustration purposes only. A person skilled in the relevant art will recognize that other components and configurations may be used without parting from the spirit and scope of the invention.

[0027] The present invention relates to reservations of resources within the context of a compute environment. One example of a compute environment is a cluster. The cluster may be, for example, a group of computing devices operated by a hosting facility, a hosting center, a virtual hosting center, a data center, grid and/or utility-based computing environments. Every reservation consists of three major components: a set of resources, a timeframe, and an access control list (ACL). Additionally, a reservation may also have a number of optional attributes controlling its behavior and interaction with other aspects of scheduling. A reservation's ACL specifies which jobs can use the reservation. Only jobs which meet one or more of a reservation's access criteria are allowed to use the reserved resources during the reservation timeframe. The reservation access criteria comprises, in one example, at least following: users, groups, accounts, classes, quality of service (QOS) and job duration. A job may be any venue or end of consumption of resource for any broad purpose, whether it be for a batch system, direct volume access or other service provisioning.

[0028] A workload manager, or scheduler, will govern access to the compute environment by receiving requests for reservations of resources and creating reservations for processing jobs. A workload manager functions by manipulating five primary, elementary objects. These are jobs, nodes, reservations, QOS structures, and policies. In addition to these, multiple minor elementary

objects and composite objects are also utilized. These objects are also defined in a scheduling dictionary.

[0029] A workload manager may operate on a single computing device or multiple computing devices to manage the workload of a compute environment. The "system" embodiment of the invention may comprise a computing device that includes the necessary hardware and software components to enable a workload manager or a software module performing the steps of the invention. Such a computing device may include such known hardware elements as one or more central processors, random access memory (RAM), read-only memory (ROM), storage devices such as hard disks, communication means such as a modem or a card to enable networking with other computing devices, a bus that provides data transmission between various hardware components, a keyboard, a display, an operating system and so forth. There is no restriction that the particular system embodiment of the invention have any specific hardware components and any known or future developed hardware configurations are contemplated as within the scope of the invention when the computing device operates as is claimed.

[0030] Job information is provided to the workload manager scheduler from a resource manager such as Loadleveler, the Portable Batch System (PBS), Wiki or Platform's LSF products. Those of skill in the art will be familiar with each of these software products and their variations. Job attributes include ownership of the job, job state, amount and type of resources required by the job, required criteria (I need this job finished in 1 hour), preferred criteria (I would like this job to complete in ½ hour) and a wallclock limit, indicating how long the resources are required. A job consists of one or more requirements each of which requests a number of resources of a given type. For example, a job may consist of two requirements, the first asking for '1 IBM node with at least 512 MB of RAM' and the second asking for '24 IBM nodes with at least 128 MB of RAM'. Each requirement consists of one or more tasks where a task is defined as the minimal independent unit of resources. A task is a collection of elementary resources which must be allocated together within a single node. For example, a task may consist of one processor, 512MB or memory, and 2 GB of local disk. A task may also be just a single processor. In symmetric multiprocessor (SMP) environments, however, users may wish to tie one or more processors together with a certain amount of memory and/or other resources. A key aspect of a task is that the resources associated with the task must be allocated as an atomic unit, without spanning node boundaries. A task requesting 2 processors cannot be satisfied by allocating 2 uni-processor nodes, nor can a task requesting 1 processor and 1 GB of memory be satisfied by allocating 1 processor on one node and memory on another.

[0031] A job requirement (or req) consists of a request for a single type of resources. Each requirement consists of the following components: (1) a task definition is a specification of the

elementary resources which compose an individual task; (2) resource constraints provide a specification of conditions which must be met in order for resource matching to occur. Only resources from nodes which meet all resource constraints may be allocated to the job requirement; (3) a task count relates to the number of task instances required by the requirement; (4) a task List is a list of nodes on which the task instances have been located; and (5) requirement statistics are statistics tracking resource utilization.

[0032] As far as the workload manager is concerned, a node is a collection of resources with a particular set of associated attributes. In most cases, it fits nicely with the canonical world view of a node such as a PC cluster node or an SP node. In these cases, a node is defined as one or more CPU's, memory, and possibly other compute resources such as local disk, swap, network adapters, software licenses, etc. Additionally, this node will described by various attributes such as an architecture type or operating system. Nodes range in size from small uni-processor PC's to large SMP systems where a single node may consist of hundreds of CPU's and massive amounts of memory.

[0033] Information about nodes is provided to the scheduler chiefly by the resource manager. Attributes include node state, configured and available resources (i.e., processors, memory, swap, etc.), run classes supported, etc.

[0034] Policies are generally specified via a configuration file and serve to control how and when jobs start. Policies include, but are not limited to, job prioritization, fairness policies, fairshare configuration policies, and scheduling policies. Jobs, nodes, and reservations all deal with the abstract concept of a resource. A resource in the workload manager world is one of the following: (1) processors which are specified with a simple count value; (2) memory such as real memory or 'RAM' is specified in megabytes (MB); (3) swap which is virtual memory or 'swap' is specified in megabytes (MB); and (4) disk space such as a local disk is specified in megabytes (MB) or gigabytes (GB). In addition to these elementary resource types, there are two higher level resource concepts used within workload manager. These are the task and the processor equivalent (PE).

[0035] In a workload manager, jobs or reservations that request resources make such a request in terms of tasks typically using a task count and a task definition. By default, a task maps directly to a single processor within a job and maps to a full node within reservations. In all cases, this default definition can be overridden by specifying a new task definition. Within both jobs and reservations, depending on task definition, it is possible to have multiple tasks from the same job mapped to the same node. For example, a job requesting 4 tasks using the default task definition of 1 processor per task, can be satisfied by two dual processor nodes.

[0036] The concept of the PE arose out of the need to translate multi-resource consumption requests into a scalar value. It is not an elementary resource, but rather, a derived resource metric.

It is a measure of the actual impact of a set of requested resources by a job on the total resources available system wide. It is calculated as:

PE = MAX(ProcsRequestedByJob / TotalConfiguredProcs,
MemoryRequestedByJob / TotalConfiguredMemory,
DiskRequestedByJob / TotalConfiguredDisk,
SwapRequestedByJob / TotalConfiguredSwap) * TotalConfiguredProcs

[0037] For example, say a job requested 20% of the total processors and 50% of the total memory of a 128 processor MPP system. Only two such jobs could be supported by this system. The job is essentially using 50% of all available resources since the system can only be scheduled to its most constrained resource, in this case memory. The processor equivalents for this job should be 50% of the PE = 64.

[0038] A further example will be instructive. Assume a homogeneous 100 node system with 4 processors and 1 GB of memory per node. A job is submitted requesting 2 processors and 768 MB of memory. The PE for this job would be calculated as:

PE = MAX(2/(100*4), 768/(100*1024)) * (100*4) = 3.

[0039] This result makes sense since the job would be consuming 3/4 of the memory on a 4 processor node. The calculation works equally well on homogeneous or heterogeneous systems, uni-processor or large way SMP systems.

[0040] A class (or queue) is a logical container object which can be used to implicitly or explicitly apply policies to jobs. In most cases, a class is defined and configured within the resource manager and associated with one or more of the attributes or constraints shown in Table 1 below.

Table 1 - Attributes of a Class

Attribute	Description
Default Job Attributes	A queue may be associated with a default job duration, default size, or default resource requirements
Host Constraints	A queue may constrain job execution to a particular set of hosts
Job Constraints	A queue may constrain the attributes of jobs which may submitted including setting limits such as max wallclock time, minimum number of processors, etc.
Access List	A queue may constrain who may submit jobs into it based on user lists, group lists, etc.
Special Access	A queue may associate special privileges with jobs including adjusted job priority.

[0041] As stated previously, most resource managers allow full class configuration within the resource manager. Where additional class configuration is required, the CLASSCFG parameter may be used. The workload manager tracks class usage as a consumable resource allowing sites to limit

the number of jobs using a particular class. This is done by monitoring class initiators which may be considered to be a ticket to run in a particular class. Any compute node may simultaneously support several types of classes and any number of initiators of each type. By default, nodes will have a one-to-one mapping between class initiators and configured processors. For every job task run on the node, one class initiator of the appropriate type is consumed. For example, a three processor job submitted to the class batch will consume three batch class initiators on the nodes where it is run.

[0042] Using queues as consumable resources allows sites to specify various policies by adjusting the class initiator to node mapping. For example, a site running serial jobs may want to allow a particular 8 processor node to run any combination of batch and special jobs subject to the following constraints:

- only 8 jobs of any type allowed simultaneously
- no more than 4 special jobs allowed simultaneously

[0043] To enable this policy, the site may set the node's MAXJOB policy to 8 and configure the node with 4 special class initiators and 8 batch class initiators. Note that in virtually all cases jobs have a one-to-one correspondence between processors requested and class initiators required. However, this is not a requirement and, with special configuration sites may choose to associate job tasks with arbitrary combinations of class initiator requirements.

[0044] In displaying class initiator status, workload manager signifies the type and number of class initiators available using the format [<CLASSNAME>:<CLASSCOUNT>]. This is most commonly seen in the output of node status commands indicating the number of configured and available class initiators, or in job status commands when displaying class initiator requirements.

[0045] Nodes can also be configured to support various arbitrary resources. Information about such resources can be specified using the NODECFG parameter. For example, a node may be configured to have "256 MB RAM, 4 processors, 1 GB Swap, and 2 tape drives".

[0046] We next turn to the concept of reservations. There are several types of reservations which sites typically deal with. The first, administrative reservations, are typically one-time reservations created for special purposes and projects. These reservations are created using a command that sets a reservation. These reservations provide an integrated mechanism to allow graceful management of unexpected system maintenance, temporary projects, and time critical demonstrations. This command allows an administrator to select a particular set of resources or just specify the quantity of resources needed. For example, an administrator could use a regular expression to request a reservation be created on the nodes 'blue0[1-9]' or could simply request that the reservation locate the needed resources by specifying a quantity based request such as 'TASKS==20'.

[0047] Another type of reservation is called a standing reservation. This is shown in FIG. 2A. A standing reservation is useful for recurring needs for a particular type of resource distribution. For example, a site could use a standing reservation to reserve a subset of its compute resources for quick turnaround jobs during business hours on Monday thru Friday. Standing reservations are created and configured by specifying parameters in a configuration file.

[0048] As shown in FIG. 2A, the compute environment 202 includes standing reservations shown as 204A, 204B and 204C. These reservations show resources allocated and reserved on a periodic basis. These are, for example, consuming reservations meaning that cluster resources will be consumed by the reservation. These reservations are specific to a user or a group of users and allow the reserved resources to be also customized specific to the workload submitted by these users or groups. For example, one aspect of the invention is that a user may have access to reservation 204A and not only submit jobs to the reserved resources but request, perhaps for optimization or to meet preferred criteria as opposed to required criteria, that the resources within the reservation be modified by virtual partitioning or some other means to accommodate the particular submitted job. In this regard, this embodiment of the invention enables the user to submit and perhaps request modification or optimization within the reserved resources for that particular job. There may be an extra charge or debit of an account of credits for the modification of the reserved resources. The modification of resources within the reservation according to the particular job may also be performed based on a number of factors discussed herein, such as criteria, class, quality of service, policies etc.

[0049] Standing reservations build upon the capabilities of advance reservations to enable a site to enforce advanced usage policies in an efficient manner. Standing reservations provide a superset of the capabilities typically found in a batch queuing system's class or queue architecture. For example, queues can be used to allow only particular types of jobs access to certain compute resources. Also, some batch systems allow these queues to be configured so that they only allow this access during certain times of the day or week. Standing reservations allow these same capabilities but with greater flexibility and efficiency than is typically found in a normal queue management system.

[0050] Standing Reservations provide a mechanism by which a site can dedicate a particular block of resources for a special use on a regular daily or weekly basis. For example, node X could be dedicated to running jobs only from users in the accounting group every Friday from 4 to 10 PM. A standing reservation is a powerful means of controlling access to resources and controlling turnaround of jobs.

[0051] Another embodiment of reservation is something called a reservation mask, which allows a site to create "sandboxes" in which other guarantees can be made. The most common aspects of this reservation are for grid environments and personal reservation environments. In a grid

environment, a remote entity will be requesting resources and will want to use these resources on an autonomous cluster for the autonomous cluster to participate. In many cases it will want to constrain when and where the entities can reserve or utilize resources. One way of doing that is via the reservation mask.

[0052] FIG. 2B illustrates the reservation mask shown as creating sandboxes 206A, 206B, 206C in compute environment 210 and allows the autonomous cluster to state that only a specific subset of resources can be used by these remote requesters during a specific subset of times. When a requester asks for resources, the scheduler will only report and return resources available within this reservation, after which point the remote entity desires it, it can actually make a consumption reservation and that reservation is guaranteed to be within the reservation mask space. The consumption reservations 212A, 212B, 212C, 212D are shown within the reservation masks.

[0053] Another concept related to reservations is the personal reservation and/or the personal reservation mask. In compute environment 210, the reservation masks operate differently from consuming reservations in that they are enabled to allow personal reservations to be created within the space that is reserved. ACL's are independent inside of a sandbox reservation or a reservation mask in that you can also exclude other requesters out of those spaces so they're dedicated for these particular users.

[0054] One benefit of the personal reservation approach includes preventing local job starvation, and providing a high level of control to the cluster manager in that he or she can determine exactly when, where, how much and who can use these resources even though he doesn't necessarily know who the requesters are or the combination or quantity of resources they will request. The administrator can determine when, how and where requestors will participate in these clusters or grids. A valuable use is in the space of personal reservations which typically involves a local user given the authority to reserve a block of resources for a rigid time frame. Again, with a personal reservation mask, the requests are limited to only allow resource reservation within the mask time frame and mask resource set, providing again the administrator the ability to constrain exactly when and exactly where and exactly how much of resources individual users can reserve for a rigid time frame. The individual user is not known ahead of time but it is known to the system, it is a standard local cluster user.

[0055] The reservation masks 206A, 206B and 206C define periodic, personal reservation masks where other reservations in the compute environment 210 may be created, i.e., outside the defined boxes. These are provisioning or policy-based reservations in contrast to consuming reservations. In this regard, the resources in this type of reservation are not specifically allocated but the time and space defined by the reservation mask cannot be reserved for other jobs. Reservation masks enable the system to be able to control the fact that resources are available for specific purposes, during

specific time frames. The time frames may be either single time frames or repeating time frames to dedicate the resources to meet project needs, policies, guarantees of service, administrative needs, demonstration needs, etc. This type of reservation insures that reservations are managed and scheduled in time as well as space. Boxes 208A, 208B, 208C and 208D represent non-personal reservation masks. They have the freedom to be placed anywhere in cluster including overlapping some or all of the reservation masks 206A, 206B, 206C. Overlapping is allowed when the personal reservation mask was setup with a global ACL. To prevent the possibility of an overlap of a reservation mask by a non-personal reservation, the administrator can set an ACL to constrain it is so that only personal consumption reservations are inside. These personal consumption reservations are shown as boxes 212B, 212A, 212C, 212D which are constrained to be within the personal reservation masks 206A, 206B, 206C. The 208A, 208B, 208C and 208D reservations, if allowed, can go anywhere within the cluster 210 including overlapping the other personal reservation masks. The result is the creation of a "sandbox" where only personal reservations can go without in any way constraining the behavior of the scheduler to schedule other requests.

[0056] All reservations possess a start and an end time which define the reservation's active time. During this active time, the resources within the reservation may only be used as specified by the reservation ACL. This active time may be specified as either a start/end pair or a start/duration pair. Reservations exist and are visible from the time they are created until the active time ends at which point they are automatically removed.

[0057] For a reservation to be useful, it must be able to limit who or what can access the resources it has reserved. This is handled by way of an access control list, or ACL. With reservations, ACL's can be based on credentials, resources requested, or performance metrics. In particular, with a standing reservation, the attributes userlist, grouplist, accountlist, classlist, qoslist, jobattrlist, proclimit, timelimit and others may be specified.

[0058] FIG. 3 illustrates an aspect of the present invention that allows the ACL 306 for the reservation 304 to have a dynamic aspect instead of simply being based on who the requester is. The ACL decision-making process is based at least in part on the current level of service or response time that is being delivered to the requester. To illustrate the operation of the ACL 306, assume that a user 308 submits a job 314 to a queue 310 and that the ACL 306 reports that the only job that can access these resources 302 are those that have a queue time that currently exceeds two hours. The resources 302 are shown with resources N on the y axis and time on the x axis. If the job 314 has sat in the queue 310 for two hours it will then access the additional resources to prevent the queue time for the user 308 from increasing significantly beyond this time frame. The decision to allocate these additional resources can be keyed off of utilization of an expansion factor and

other performance metrics of the job. For example, the reservation 304 may be expanded or contracted or migrated to cover a new set of resources.

[0059] Whether or not an ACL 306 is satisfied is typically and preferably determined the scheduler 104A. However, there is no restriction in the principle of the invention regarding where or on what node in the network the process of making these allocation of resource decisions occurs. The scheduler 104A is able to monitor all aspects of the request by looking at the current job 314 inside the queue 310 and how long it has sat there and what the response time target is and the scheduler itself determines whether all requirements of the ACL 306 are satisfied. If requirements are satisfied, it releases the resources that are available to the job 314. A job 314 that is located in the queue and the scheduler communicating with the scheduler 104A. If resources are allocated, the job 314 is taken from the queue 310 and inserted into the reservation 314 in the cluster 302.

[0060] An example benefit of this model is that it makes it significantly easier for a site to balance or provide guaranteed levels of service or constant levels of service for key players or the general populace. By setting aside certain resources and only making them available to the jobs which threaten to violate their quality of service targets, the system increases the probability of satisfying targets.

[0061] When specifying which resources to reserve, the administrator has a number of options. These options allow control over how many resources are reserved and where they are reserved at. The following reservation attributes allow the administrator to define resources.

[0062] An important aspect of reservations is the idea of a task. The scheduler uses the task concept extensively for its job and reservation management. A task is simply an atomic collection of resources, such as processors, memory, or local disk, which must be found on the same node. For example, if a task requires 4 processors and 2 GB of memory, the scheduler must find all processors AND memory on the same node; it cannot allocate 3 processors and 1 GB on one node and 1 processor and 1 GB of memory on another node to satisfy this task. Tasks constrain how the scheduler must collect resources for use in a standing reservation, however, they do not constrain the way in which the scheduler makes these cumulative resources available to jobs. A job can use the resources covered by an accessible reservation in whatever way it needs. If reservation X allocated 6 tasks with 2 processors and 512 MB of memory each, it could support job Y which requires 10 tasks of 1 processor and 128 MB of memory or job Z which requires 2 tasks of 4 processors and 1 GB of memory each. The task constraints used to acquire a reservation's resources are completely transparent to a job requesting use of these resources. Using the task description, the taskcount attribute defines how many tasks must be allocated to satisfy the reservation request. To create a reservation, a taskcount and/or a hostlist may be specified.

[0063] A hostlist constrains the set of resource which are available to a reservation. If no taskcount is specified, the reservation will attempt to reserve one task on each of the listed resources. If a taskcount is specified which requests fewer resources than listed in the hostlist, the scheduler will reserve only the number of tasks from the hostlist specified by the taskcount attribute. If a taskcount is specified which requests more resources than listed in the hostlist, the scheduler will reserve the hostlist nodes first and then seek additional resources outside of this list.

[0064] Reservation flags allow specification of special reservation attributes or behaviors. Supported flags are listed in table 2 below.

Table 2

Flag Name	Description
BESTEFFORT	N/A
BYNAME	reservation will only allow access to jobs which meet reservation ACL's *and* explicitly request the resources of this reservation using the job ADVRES flag
IGNRSV	request will ignore existing resource reservations allowing the reservation to be *forced* onto available resources even if this conflicts with other reservations.
OWNERPREEMPT	job's by the reservation owner are allowed to preempt non-owner jobs using reservation resources
PREEMPTEE	Preempts a job or other object
SINGLEUSE	reservation is automatically removed after completion of the first job to use the reserved resources
SPACEFLEX	reservation is allowed to adjust resources allocated over time in an attempt to optimize resource utilization
TIMEFLEX	reservation is allowed to adjust the reserved timeframe in an attempt to optimize resource utilization

[0065] Reservations must explicitly request the ability to float for optimization purposes by using a flag such as the SPACEFLEX flag. The reservations may be established and then identified as self-optimizing in either space or time. If the reservation is flagged as such, then after the reservation is created, conditions within the compute environment may be monitored to provide feedback on where optimization may occur. If so justified, a reservation may migrate to a new time or migrate to a new set of resources that are more optimal than the original reservation.

[0066] FIG. 4 illustrates a reservation creation window 400 that includes the use of the flags in Table 2. A user Scott input reservation information in a variety of fields 402 for name, partition, node features and floating reservation. Each of these input fields includes a drop-down menu to enable the selection of options easy. An access control list input field 404 allows the user to input an account, class/queue, user, group and QoS information. Resources may be assigned and searched and tasks created 406 and reservation flags set 408, such as best effort, single use,

preemptee, time flex, by name, owner preempt, space flex, exclusive and force. These flags set parameters that may cause the reservation to be optimized such as in time or space where it migrates to a new time or over new resources based on monitored events or other feedback.

[0067] A reservation time-frame 410 may also be input such as one, daily, weekly, with start and end times for the reservation. Menu drop down calendars and clocks are available for easily enabling the user to view and graphically input and select the various timeframe parameters. Event triggers may also be input wherein the user can create one or more triggers associated with the reservation. As generally shown in FIG. 4, the use of a graphical interface simplifies the reservation-creation process for the administrator or user.

[0068] FIG. 5 illustrates a particular instance where the user has identified the time-flex and space-flex flags within the reservation. A window 500 identifies three reservations 502 for 96 nodes, 504 for 128 nodes and 506 for 256 nodes. The height of each of these reservations generally relates to resources reserved, such as a number of processors reserved or processors and disk space. The X-axis represents time. Reservation 508 represents a reservation in the future that will in a position to receive submitted jobs. Assume that reservation 506 which was scheduled to end at time T2 has finished early at time T1. Also assume that reservation 508 is flagged for time flex and space flex. In this case, based on the monitored event that reservation 506 has ended early, the system would cause reservation 508 to migrate in time (and space in this example) to position 510. This represents a movement of the reservation to a new time and a new set of resources. If reservation 504 ends early, and reservation 508 migrates to position 520, that would represent a migration in time (to an earlier time) but not in space. This would be enabled by the time-flex flag being set wherein the migration would seek to create a new reservation at the earliest time possible and/or according to available resources. The new time may be based on criteria to minimize the time for the reservation or to maximize usage of the overall resources or better performance of the compute environment.

[0069] Next, assume that reservation 508 is for 128 processors and reservation 506 is for 256 processors and reservation 508 is flagged for space flex. If reservation 506 ends are time T1 instead of time T2, then reservation 508 may migrate to position 512 to a reservation of 256 processors. The time frame of the starting and ending time may be the same but the reservation has migrated in space and thus been optimized.

[0070] In another aspect of reservation migration, assume that reservation 508 is set but that a node or a group of nodes that are part of the reservation go down or are projected to fail as represented by 518. In this regard, reservation 508 may be enabled to migrate as shown by 516 and 508 to cover new resources but to accommodate for the nodes that are no longer available.

[0071] Standing reservations allow resources to be dedicated for particular uses. This dedication can be configured to be permanent or periodic, recurring at a regular time of day and/or time of week. There is extensive applicability of standing reservations for everything from daily dedicated job runs to improved use of resources on weekends. All standing reservation attributes are specified via a parameter using available attributes

[0072] In addition to standing and administrative reservations, a workload manager according to the invention can also create priority reservations. These reservations are used to allow the benefits of out-of-order execution (such as is available with a backfill feature) without the side effect of job starvation. Starvation can occur in any system where the potential exists for a job to be overlooked by the scheduler for an indefinite period. In the case of backfill, small jobs may continue to be run on available resources as they become available while a large job sits in the queue never able to find enough nodes available simultaneously to run on. To avoid such situations, priority reservations are created for high priority jobs which cannot run immediately. When making these reservations, the scheduler determines the earliest time the job could start, and then reserves these resources for use by this job at that future time. By default, only the highest priority job will receive a priority reservation. However, this behavior is configurable via a reservation depth policy. The workload manager's default behavior of only reserving the highest priority job allows backfill to be used in a form known as liberal backfill. This liberal backfill tends to maximize system utilization and minimize overall average job turnaround time. However, it does lead to the potential of some lower priority jobs being indirectly delayed and may lead to greater variance in job turnaround time. A reservation depth parameter can be set to a very large value, essentially enabling what is called conservative backfill where every job which cannot run is given a reservation. Most sites prefer the liberal backfill approach associated with the default reservation depth 1 or select a slightly higher value. It is important to note that to prevent starvation in conjunction with reservations, monotonically increasing priority factors such as queuetime or job x-factor should be enabled.

[0073] Another important consequence of backfill and reservation depth is its affect on job priority. In the workload manager, all jobs are preferably prioritized. Backfill allows jobs to be run out of order and thus, to some extent, job priority to be ignored. This effect, known as 'priority dilution' can cause many site policies implemented via workload manager prioritization policies to be ineffective. Setting the reservation depth parameter to a higher value will give job priority 'more teeth' at the cost of slightly lower system utilization. This lower utilization results from the constraints of these additional reservations, decreasing the scheduler's freedom and its ability to find additional optimizing schedules. Anecdotal evidence indicates that these utilization losses are fairly minor, rarely exceeding 8%.

[0074] In addition to the reservation depth parameter, sites also have the ability to control how reservations are maintained. The workload manager's dynamic job prioritization allows sites to prioritize jobs so that their priority order can change over time. It is possible that one job can be at the top of the priority queue for a time, and then get bypassed by another job submitted later. A reservation policy parameter allows a site to determine what how existing reservations should be handled when new reservations are made. The value "highest" will cause that all jobs which have ever received a priority reservation will maintain that reservation until they run even if other jobs later bypass them in priority value. The value of the parameter "current highest" will cause that only the current top <RESERVATIONDEPTH> priority jobs will receive reservations. If a job had a reservation but has been bypassed in priority by another job so that it no longer qualifies as being among the top <RESERVATIONDEPTH> jobs, it will lose its reservation. Finally, the value "never" indicates that no priority reservations will be made.

[0075] QOS-based reservation depths can be enabled via the reservation QOS list parameter. This parameter allows varying reservation depths to be associated with different sets of job QoS's. For example, the following configuration will create two reservation depth groupings:

```
---
RESERVATIONDEPTH[0]  8
RESERVATIONQOSLIST[0] highprio interactive debug

RESERVATIONDEPTH[1]  2
RESERVATIONQOSLIST[1] batch
---
```

[0076] This example will cause that the top 8 jobs belonging to the aggregate group of highprio, interactive, and debug QoS jobs will receive priority reservations. Additionally, the top 2 batch QoS jobs will also receive priority reservations. Use of this feature allows sites to maintain high throughput for important jobs by guaranteeing a significant proportion of these jobs are making progress toward starting through use of the priority reservation. The following are example default values for some of these parameters: RESERVATIONDEPTH[DEFAULT] = 1; RESERVATIONQOSLIST[DEFAULT] = ALL.

[0077] This allows one job with the highest priority to get a reservation. These values can be overwritten by modifying the default policy.

[0078] A final reservation policy is in place to handle a number of real-world issues. Occasionally when a reservation becomes active and a job attempts to start, various resource manager race conditions or corrupt state situations will prevent the job from starting. By default, the workload manager assumes the resource manager is corrupt, releases the reservation, and attempts to re-create

the reservation after a short timeout. However, in the interval between the reservation release and the re-creation timeout, other priority reservations may allocate the newly available resources, reserving them before the original reservation gets an opportunity to reallocate them. Thus, when the original job reservation is re-established, its original resource may be unavailable and the resulting new reservation may be delayed several hours from the earlier start time. The parameter reservation retry time allows a site that is experiencing frequent resource manager race conditions and/or corruption situations to tell the workload manager to hold on to the reserved resource for a period of time in an attempt to allow the resource manager to correct its state.

[0079] Next we discuss the use of partitions. Partitions are a logical construct which divide available resources and any single resource (i.e., compute node) may only belong to a single partition. Often, natural hardware or resource manager bounds delimit partitions such as in the case of disjoint networks and diverse processor configurations within a cluster. For example, a cluster may consist of 256 nodes containing four 64 port switches. This cluster may receive excellent interprocess communication speeds for parallel job tasks located within the same switch but sub-stellar performance for tasks which span switches. To handle this, the site may choose to create four partitions, allowing jobs to run within any of the four partitions but not span them.

[0080] While partitions do have value, it is important to note that within the workload manager, the standing reservation facility provides significantly improved flexibility and should be used in the vast majority of politically motivated cases where partitions may be required under other resource management systems. Standing reservations provide time flexibility, improved access control features, and more extended resource specification options. Also, another workload manager facility called node sets allows intelligent aggregation of resources to improve per job node allocation decisions. In cases where system partitioning is considered for such reasons, node sets may be able to provide a better solution.

[0081] An important aspect of partitions over standing reservations and node sets is the ability to specify partition specific policies, limits, priorities, and scheduling algorithms although this feature is rarely required. An example of this need may be a cluster consisting of 48 nodes owned by the Astronomy Department and 16 nodes owned by the Mathematics Department. Each department may be willing to allow sharing of resources but wants to specify how their partition will be used. As mentioned earlier, many of the workload manager's scheduling policies may be specified on a per partition basis allowing each department to control the scheduling goals within their partition.

[0082] The partition associated with each node should be specified as indicated in the node location section. With this done, partition access lists may be specified on a per job or per QOS basis to constrain which resources a job may have access to. By default, QOS's and jobs allow

global partition access. Note that by default, a job may only utilize resources within a single partition.

[0083] If no partition is specified, the workload manager creates one partition per resource manager into which all resources corresponding to that resource manager are placed. This partition may be given the same name as the resource manager. A partition preferably does not span multiple resource managers. In addition to these resource manager partitions, a pseudo-partition named [ALL] is created which contains the aggregate resources of all partitions. While the resource manager partitions are real partitions containing resources not explicitly assigned to other partitions, the [ALL] partition is only a convenience object and is not a real partition; thus it cannot be requested by jobs or included in configuration ACL's.

[0084] Node-to-partition mappings are established using a node configuration parameter as shown in this example:

```
NODECFG[node001]  PARTITION=astronomy
NODECFG[node002]  PARTITION=astronomy
...
NODECFG[node049]  PARTITION=math
...
```

[0085] By default, the workload manager only allows the creation of 4 partitions total. Two of these partitions, DEFAULT, and [ALL], are used internally, leaving only two additional partition definition slots available. If more partitions will be needed, the maximum partition count should be adjusted. Increasing the maximum number of partitions can be managed.

[0086] Determining who can use which partition is specified using *CFG parameters (for example, these parameters may be defined as: usercfg, groupcfg, accountcfg, quoscfg, classcfg and systemcfg). These parameters allow both a partition access list and default partition to be selected on a credential or system wide basis using the PLIST and PDEF keywords. By default, the access associated with any given job is the logical or of all partition access lists assigned to the job's credentials. Assume a site with two partitions: general and test. The site management would like everybody to use the general partition by default. However, one user, Steve, needs to perform the majority of his work on the test partition. Two special groups, staff and mgmt will also need access to use the test partition from time to time but will perform most of their work in the general partition. The example configuration below will enable the needed user and group access and defaults for this site.

```
SYSCFG[base]    PLIST=
USERCFG[DEFAULT] PLIST=general
USERCFG[steve]  PLIST=general:test PDEF=test
```

GROUPCFG[staff] PLIST=general:test PDEF=general

GROUPCFG[mgmt] PLIST=general:test PDEF=general

[0087] By default, the system partition access list allows global access to all partitions. If using logically or based partition access lists, the system partition list should be explicitly constrained using the SYSCFG parameter. While using a logical or approach allows sites to add access to certain jobs, some sites prefer to work the other way around. In these cases, access is granted by default and certain credentials are then restricted from access various partitions. To use this model, a system partition list must be specified. See the example below:

SYSCFG[base] PLIST=general,test&

USERCFG[demo] PLIST=test&

GROUPCFG[staff] PLIST=general&

[0088] In the above example, note the ampersand ('&'). This character, which can be located anywhere in the PLIST line, indicates that the specified partition list should be logically AND'd with other partition access lists. In this case, the configuration will limit jobs from user demo to running in partition test and jobs from group staff to running in partition general. All other jobs will be allowed to run in either partition. When using and based partition access lists, the base system access list must be specified with SYSCFG.

[0089] Users may request to use any partition they have access to on a per job basis. This is accomplished using the resource manager extensions, since most native batch systems do not support the partition concept. For example, on a PBS system, a job submitted by a member of the group staff could request that the job run in the test partition by adding the line '#PBS -W x=PARTITION:test' to the command file. Special jobs may be allowed to span the resources of multiple partitions if desired by associating the job with a QOS which has the flag 'SPAN' set.

[0090] The disclosure now continues to discuss reservations further. An advance reservation is the mechanism by which the present invention guarantees the availability of a set of resources at a particular time. With an advanced reservation a site now has an ability to actually specify how the scheduler should manage resources in both space and time. Every reservation consists of three major components, a list of resources, a timeframe (a start and an end time during which it is active), and the ACL. These elements are subject to a set of rules. The ACL acts as a doorway determining who or what can actually utilize the resources of the cluster. It is the job of the cluster scheduler to make certain that the ACL is not violated during the reservation's lifetime (i.e., its timeframe) on the resources listed. The ACL governs access by the various users to the resources. The ACL does this by determining which of the jobs, various groups, accounts, jobs with special service levels, jobs with requests for specific resource types or attributes and many different aspects of requests can actually come in and utilize the resources. With the ability to say that these

resources are reserved, the scheduler can then enforce true guarantees and can enforce policies and enable dynamic administrative tasks to occur. The system greatly increases in efficiency because there is no need to partition the resources as was previously necessary and the administrative overhead is reduced it terms of staff time because things can be automated and scheduled ahead of time and reserved.

[0091] As an example of a reservation, a reservation may specify that node002 is reserved for user John Doe on Friday. The scheduler will thus be constrained to make certain that only John Doe's jobs can use node002 at any time on Friday. Advance reservation technology enables many features including backfill, deadline based scheduling, QOS support, and meta scheduling.

[0092] There are several reservation concepts that will be introduced as aspects of the invention. These include dynamic reservations, co-allocating reservation resources of different types, reservations that self-optimize in time, reservations that self-optimization in space, reservations rollbacks and reservation masks. The present invention relates to a system and method of providing dynamic reservations in a compute environment. Dynamic reservations are reservations that are able to be modified once they are created. The workload manager allows dynamic modification of most scheduling parameters allowing new scheduling policies, algorithms, constraints, and permissions to be set at any time. For example, a reservation may be expanded or contracted after a job is submitted to more closely match the reservation to the workload. Changes made via client commands are preferably temporary and will be overridden by values specified in a config files the next time the workload manager is shutdown and restarted.

[0093] Various commands may be used manually or automatically to control reservations. Examples of such commands and their function are illustrated in Table 3:

Table 3

mdiag	-r	display summarized reservation information and any unexpected state
mrsvctl		reservation control
mrsvctl	-r	remove reservations
mrsvctl	-c	create an administrative reservation
showres		display information regarding location and state of reservations

[0094] FIG. 6 illustrates a method of dynamically modifying a request, a reservation or the compute environment. Attributes of a reservation may change based on a feedback mechanism that adds intelligence as to ideal characteristics of the reservation and how it should be applied as the context of its environment or an entities needs change. One example of a dynamic reservation is a reservation that provides for a guarantee of resources for a request unless no jobs that consume resources are submitted under the request or if the user is not using the reserved resources. In

other words, if no jobs are submitted on reserved resources or the job that is submitted does not need all of the reserved resources.

[0095] The example method in FIG. 6 can relate to the scenario where a job has or has not yet been submitted to reserved compute resources. The method comprises receiving a request for resources within the compute environment (602) and monitoring events after receiving the request for resources (604). Based on the monitored events, the method comprises dynamically modifying at least one of the request for resources, a reservation and the compute environment (606). The compute environment may be a computer farm, a cluster, a grid, an on-demand computing center and the like.

[0096] The request for resources may be a request for consumption of resources such as processor time and network bandwidth. The request may also be for provisioning resources such as available licenses for particular software or operating systems. The request may also be for such things as a request to process a batch job or for direct volume access, or a request for a virtual private cluster.

[0097] The monitored events may further mean monitoring events related to the compute environment. Events that may be identified include, but are not limited to, new resources that become available because other jobs finish early, compute nodes that go down and are unavailable, other jobs submitted to the compute environment. In this regard, the monitoring may include jobs submitted by an administrator, other users or the requestor. For example, if the requestor never submits a job within a reservation made according to the request, then the method may modify the reservation by shrinking the reservation or reduce the reserved amount of resources for efficiency. The request or the reservation may also be canceled if no jobs are submitted or based on other criteria.

[0098] A job submitted may also be one of a reservation, an object that monitors policy, an object that monitors credentials, an object that monitors node states and an object that monitors the compute environment. If the compute environment is dynamically modified according to the monitored events, the modification may be performed to satisfy the request for resources or preferences within the request. The modifications to the compute environment may also be constrained within the reservation space.

[0099] Examples of modifications that may be done to the compute environment include but are not limited to modifying a node or nodes, modifying at least one operating system or other software, installing end-user applications, dynamically partitioning node resources and adjusting network configurations. Once a job has been submitted, the compute resources may be dynamically modified to more adequately process the job or more efficiently process the job. For example, if it is foreseen that the job will end early, the system may shorten the reservation of time for the resources to free-up migration of other reservations in that time and space. Another example may

exist where if a reservation is partly consumed by a job, but based on monitored events, the remaining reserved resources, say 128 nodes, could be expanded to 256 nodes such that the job may finish early. In that case, the reservation from the current time would be dynamically modified to include additional resources. These changes may be based on

[00100] The modifications to a request, a reservation or a compute environment may be based on a policy. For example, a dynamic reservation policy may apply which says that if the project does not use more than 25% of what it is guaranteed by the time that 50% of its time has expired, then, based on the feedback, the system dynamically modifies the reservation of resources to more closely match the job (606). In other words, the reservation dynamically adjust itself to reserve X% fewer resources for this project, thus freeing up unused resource for others to use.

[00101] If the party submitting the request for resources has not submitted a job for processing after a predetermined amount of time, then the request for resources or the job submitted to the reservation may be cancelled. This is illustrated more with reference to FIG. 7 which illustrates this reservation. A self-terminating reservation is a reservation that can cancel itself if certain criteria take place. A reservation of compute resources is created (702) and the system monitors events associated with the reservation (704). The system determines whether the monitored events justifies canceling the reservation or jobs submitted according to the reservation (706). If no, there is no justification to terminate, then the system continues to monitor events in step (704). If, however, conditions justify terminating one of the reservation or a job, then the reservation terminates itself or a job is cancelled (708).

[00102] An example of a self-terminating reservation is a reservation that uses an event policy to check that if after 30 minutes no jobs have been submitted against the reservation, or if utilization of the assigned resources is below x% then the reservation will cancel itself, thus making those resources available to be used by others. Another example is if a job is submitted to the reserved cluster resources, but to process the job would require the use of compute resources beyond the reservation time or the reserved cluster resources, then the job may be canceled and notification provided to the submitted regarding the reasons for the cancellation. Options may then be provided to the submitter for modifying the reservation, or modifying the job and so forth to enable the job to be resubmitted under modified circumstances that may enable the job to be processed.

[00103] Based on the monitored events in the cluster environment, modifying the request for resources may involve dynamically modifying the compute environment or modifying the compute environment to more adequately process jobs submitted within the reservation.

[00104] Preferably, the option of extending the reservation to accommodate the job is subject to pre-established policies that are either required or preferred. One example of presenting these types of offers includes presenting the submitter the option of extending the reservation according to a

pricing plan that would meet the preferred policies. This pricing plan may include options to pay for extended time, extended or modified resources, licenses, other provisioning options and so forth. Any combination of job or resource modification is envisioned. In this regard, the reservation of resources could migrate to "cover" a new set of resources that may meet a preferred criteria, an increased payment plan, or some other threshold. The migration of a reservation may be in both space (compute resources) and time (such as, for example, to move the start time of the reservation to as soon as possible). The migration in space may be for the purpose of increasing the performance for the overall compute environment or may be for optimizing the time of completion for a job or jobs. The migration may be for any other reason such as to modify the resources used because of a node failure or a projected maintenance of other failure of a resource. The system may also present a user with the option of allowing jobs running within a personal reservation to complete although the job is projected to run beyond the window of time for the reservation of resources.

[00105] As mentioned above, the option of extending or modifying a reservation may be based on pre-established policies that govern whether a reservation may be modified and to what extent it may be modified. There are preferably thresholds established in time and space governing the modifications.

[00106] The request for resources in a compute environment may include a request for a reservation of resources for a window of time in which at least one user can submit personal reservations. A personal reservation is a non-administrator reservation submitted by an individual user or a group of users that are not considered administrators. The personal reservation may be submitted by an administrator but is of a non-administrative nature. The window of time may also be a request for cluster resources for a periodic window of time, such as daily, weekly, monthly, quarterly and so on. Then, if the system receives a personal reservation for the use of compute resources within the window of time, the system provides access to the reserved cluster resources for the personal reservation to process submitted jobs. If the processing personal reservation exceeds the window of time for the reservation of compute resources, then the system may cancel and lock out the personal reservation from access to the cluster resources. Before canceling and locking out the personal reservation, the system may present to a user who submitted the personal reservation an option of allowing the personal reservation to complete although it is beyond the window of time for their reservation of compute resources. If a job submitted under the personal reservation would exceed the personal reservation, then the system may extend the personal reservation to meet the needs of the job or perform some other modification. A consumption job submitted may exceed the window of time allowed for the reservation and thus the system may never start the consumption job in the first place.

[00107] Charging for resource use and reservation is also an aspect of the present invention. The system may also charge the requestor a specific rate for reserved resources and a different rate for consumed resources. Yet a different rate may be charged for reserved resources that are never used.

[00108] The user/requestor may be charged for use of the resources in the cluster environment in a variety of ways. For example, the user may be charged for reserved resources at one rate, and another rate for reserved and consumed resources.

[00109] Within a reservation, the system may provide a modification of the compute resources within the reservation space. For example, the system may optimize the use of resources within that reservation to meet needs and preferences of particular jobs submitted under that reservation.

[00110] Another dynamic reservation may perform the following step: if usage of resources provided by a reservation is above 90% with fewer than 10 minutes left in the reservation then the reservation will attempt to add 10% more time to the end of the reservation to help ensure the project is able to complete. In summary, it is the ability for a reservation to receive manual or automatic feedback to an existing reservation in order to have it more accurately match any given needs, whether those be of the submitting entity, the community of users, administrators, etc. The dynamic reservation improves the state of the art by allowing the ACL to the reservation to have a dynamic aspect instead of simply being based on who the requestor is. The reservation can be based on a current level of service or response time being delivered to the requestor.

[00111] The ACL and scheduler are able to monitor all aspects of the request by looking at the current job inside the queue and how long it has sat there and what the response time target is. It is preferable, although not required, that the scheduler itself determines whether all requirements of the ACL are satisfied. If the requirements are satisfied, the scheduler releases the resources that are available to the job.

[00112] The benefits of this model is it makes it significantly easier for a site to balance or provide guaranteed levels of service or constant levels of service for key players or the general populace. By setting aside certain resources and only making them available to the jobs which threaten to violate their quality of service targets it increases the probability of satisfying it.

[00113] As can be appreciated, the methods described above for managing a compute environment provide marked improvements in how resources are reserved and how those reservations are managed in connection with the compute environment to maximize efficiency for both the user and the compute environment.

[00114] Embodiments within the scope of the present invention may also include computer-readable media for carrying or having computer-executable instructions or data structures stored thereon. Such computer-readable media can be any available media that can be accessed by a general purpose or special purpose computer. By way of example, and not limitation, such

computer-readable media can comprise RAM, ROM, EEPROM, CD-ROM or other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to carry or store desired program code means in the form of computer-executable instructions or data structures. When information is transferred or provided over a network or another communications connection (either hardwired, wireless, or combination thereof) to a computer, the computer properly views the connection as a computer-readable medium. Thus, any such connection is properly termed a computer-readable medium. Combinations of the above should also be included within the scope of the computer-readable media.

[00115] Computer-executable instructions include, for example, instructions and data which cause a general purpose computer, special purpose computer, or special purpose processing device to perform a certain function or group of functions. Computer-executable instructions also include program modules that are executed by computers in stand-alone or network environments. Generally, program modules include routines, programs, objects, components, and data structures, etc. that perform particular tasks or implement particular abstract data types. Computer-executable instructions, associated data structures, and program modules represent examples of the program code means for executing steps of the methods disclosed herein. The particular sequence of such executable instructions or associated data structures represents examples of corresponding acts for implementing the functions described in such steps.

[00116] Those of skill in the art will appreciate that other embodiments of the invention may be practiced in network computing environments with many types of computer system configurations, including personal computers, hand-held devices, multi-processor systems, microprocessor-based or programmable consumer electronics, network PCs, minicomputers, mainframe computers, and the like. Embodiments may also be practiced in distributed computing environments where tasks are performed by local and remote processing devices that are linked (either by hardwired links, wireless links, or by a combination thereof) through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

[00117] Although the above description may contain specific details, they should not be construed as limiting the claims in any way. Other configurations of the described embodiments of the invention are part of the scope of this invention. Accordingly, the appended claims and their legal equivalents should only define the invention, rather than any specific examples given.

CLAIMS

I claim:

1. A method of dynamically modifying resources within a compute environment, the method comprising:

receiving a request for resources in the compute environment;

monitoring events after receiving the request for resources; and

based on the monitored events, dynamically modifying at least one of the request for resources and the compute environment.

2. The method of claim 1, wherein the compute environment is one of a compute farm, a cluster environment and a grid environment

3. The method of claim 1, wherein the request for resources is a request for consumption resources.

4. The method of claim 1, wherein the request for resources is a request for provisioning services.

5. The method of claim 1, wherein the request for resources is a request to process a batch job.

6. The method of claim 1, wherein the request for resources is a request for direct volume access.

7. The method of claim 1, wherein the request for resources is a request for a virtual private cluster.

8. The method of claim 1, wherein monitoring events after receiving the request for resources further comprises monitoring the compute environment.

9. The method of claim 1, wherein monitoring events after receiving the request for resources further comprises monitoring to determine if a party submitting the request has submitted a job for processing once resources in the compute environment are reserved for the job.

10. The method of claim 9, wherein if the party submitting the request for resources has not submitted a job for processing after a predetermined amount of time, then dynamically modifying the request for resources further comprises canceling the request for resources.

11. The method of claim 10, wherein a job comprises one of a reservation, an object that monitors policy, an object that monitors credentials, an object that monitors node states and an object that monitors the compute environment.

12. The method of claim 11, wherein based on the monitored events in the compute environment, modifying the compute environment further comprises dynamically modifying the compute environment to satisfy the request for resources.

13. The method of claim 12, wherein dynamically modifying the compute environment further comprises at least one of: modifying at least one node, modifying at least one operating system, installing end user applications, dynamically partitioning node resources and adjusting network configuration.

14. The method of claim 1, wherein the request for resources is a request for a reservation of resources in the compute environment.

15. The method of claim 14, wherein monitoring events after receiving the request for a reservation further comprises monitoring compute resources associated with the reservation.

16. The method of claim 15, further comprising dynamically modifying the compute environment to more adequately process jobs submitted within the reservation.

17. The method of claim 1, wherein modifying the request for resources comprises migrating a reservation to be associated with new resources.

18. The method of claim 17, wherein migrating the reservation is one of a migration in space and a migration in time to the new resources.

19. The method of claim 17, wherein the new resources better meet needs associated with the request for resources.

20. The method of claim 18, wherein the migration in time seeks to create a reservation at the earliest time possible.

21. The method of claim 18, wherein the migration in time seeks to create a reservation based on availability of resources in the compute environment.

22. The method of claim 18, wherein the migration in space comprises migrating the reservation to resources that will provide better performance of the compute environment for the request for resources.

23. The method of claim 18, wherein the migration in space comprises migrating the reservation to resources according to a failure or projected failure of resources.

24. The method of claim 1, wherein monitoring events after receiving the request for resources further comprises monitoring a job submitted within a reservation based on the request.

25. The method of claim 24, wherein if the job submitted within the reservation will extend beyond the reservation, the method further comprises canceling the job.

26. The method of claim 25, wherein prior to canceling the job, the method further comprises presenting to the entity that submitted the job an option of extending the reservation to accommodate the job.

27. The method of claim 26, wherein the option of extending the reservation to accommodate the job is subject to pre-established policies.

28. The method of claim 27, further comprising presenting to the entity, with the option of extending the reservation, a pricing option to extend the reservation.

29. The method of claim 1, wherein the request for resources in a compute environment comprises a reservation of resources for a window of time in which at least one user can submit personal reservations.

30. The method of claim 29, wherein personal reservations are one of a non-administrator reservation and an administrator reservation.

31. The method of claim 29, wherein the reservation of compute resources for a window of time is a request for cluster resources for a periodic window of time.

32. The method of claim 31, wherein the periodic window of time may be daily, weekly, monthly, quarterly or yearly.

33. The method of claim 29, further comprising:

receiving a personal reservation for the use of compute resources within the window of time; and

providing access to the reserved compute resources for the personal reservation to process jobs.

34. The method of claim 33, wherein if a received consumption job associated with the personal reservation will exceed the window of time for the reservation of compute resources, then the method comprises canceling and locking out the personal reservation from access to the compute resources.

35. The method of claim 33, wherein if a received consumption job associated with the personal reservation will exceed the window of time, then the method comprises never starting the consumption job.

36. The method of claim 34, further comprising, before canceling and locking out the personal reservation, the step of:

presenting to a user who submitted the personal reservation an option of allowing the jobs running within the personal reservation to complete although it is beyond the window of time for their reservation of compute resources.

37. The method of claim 34, further comprising, if the job submitted under a personal reservation would exceed the personal reservation, extending the personal reservation to meet the needs of the job.

38. A method of managing resources within a compute environment, the method comprising:

receiving a request for resources in the compute environment;

reserving resources in the compute environment according to the request; and

charging the requestor based on the reservation of resources.

39. The method of claim 38, wherein charging the requestor further comprises charging a specific rate for the reserved resources whether the reserved resources are used or not.

40. The method of claim 38, wherein charging the requestor further comprises charging a first rate for reserved resources that are used and a second rate for reserved resources that are not used.

41. The method of claim 40, wherein a used resources is consumed by a job run within the reservation.

42. The method of claim 1, further comprising:

creating a reservation for resources within the compute environment based on the request for resource; and

dynamically customizing resources within the reservation to meet workload submitted within the reservation.

43. The method of claim 42, wherein the reservation is associated with one of an individual or a group.

44. A computer-readable medium storing instructions for controlling a computing device to dynamically manage resources within a compute environment, the instructions comprising:

receiving a request for resources in the compute environment;

monitoring events after receiving the request for resources; and

based on the monitored events, dynamically modifying at least one of the request for resources and the compute environment.

45. A system for dynamically managing resources within a compute environment, the system comprising:

means for receiving a request for resources in the compute environment;

means for monitoring events after receiving the request for resources; and

based on the monitored events, means for dynamically modifying at least one of the request for resources and the compute environment.

46. A system for dynamically managing resources within a compute environment, the system comprising:

a module configured to receive a request for resources in the compute environment;

a module configured to monitor events after receiving the request for resources; and

a module configured to dynamically modify at least one of the request for resources and the compute environment based on the monitored events.

47. A compute environment comprising a plurality of computing devices, the compute environment having resources which are dynamically managed according to a method comprising:

receiving a request for resources in the compute;

monitoring events after receiving the request for resources; and

based on the monitored events, dynamically modifying at least one of the request for resources and the compute environment.

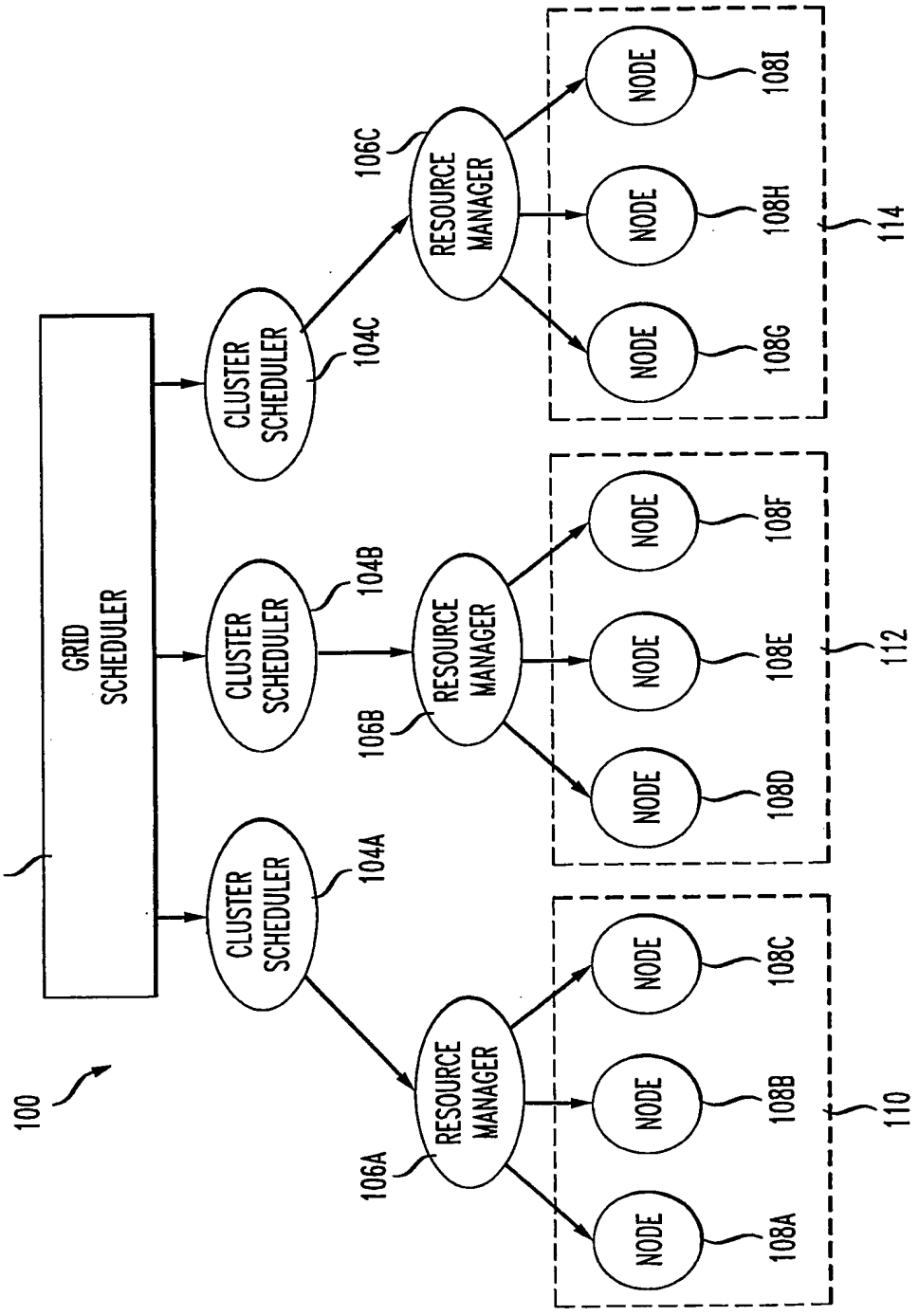


FIG. 1A
PRIOR ART

FIG. 1B
PRIOR ART

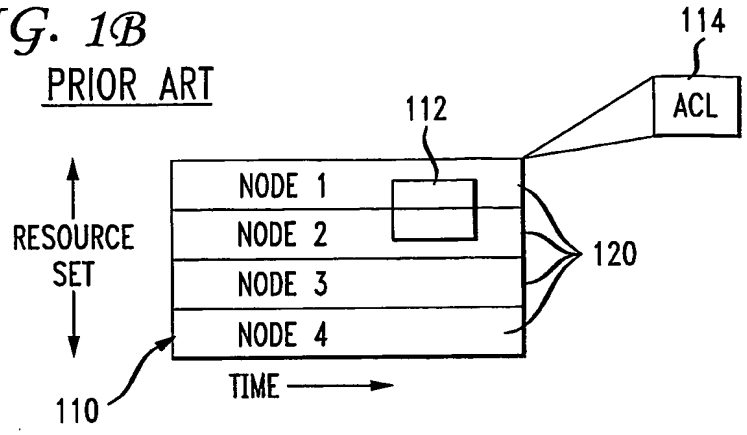


FIG. 2A

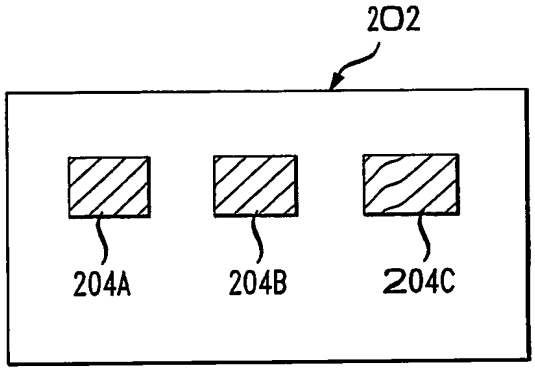


FIG. 2B

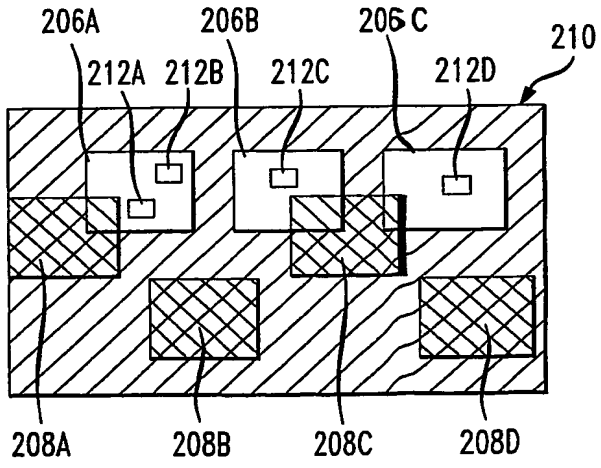


FIG. 3

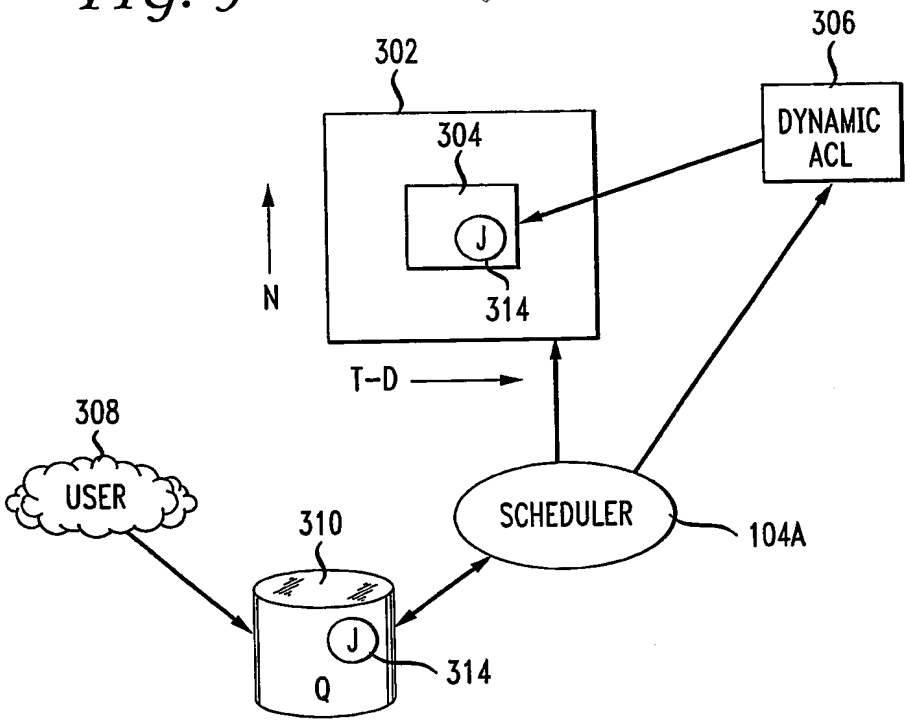

306
302
DYNAMIC ACL
304
J
314
N
T-D
308
USER
310
SCHEDULER
104A
J
314
Q

FIG. 4

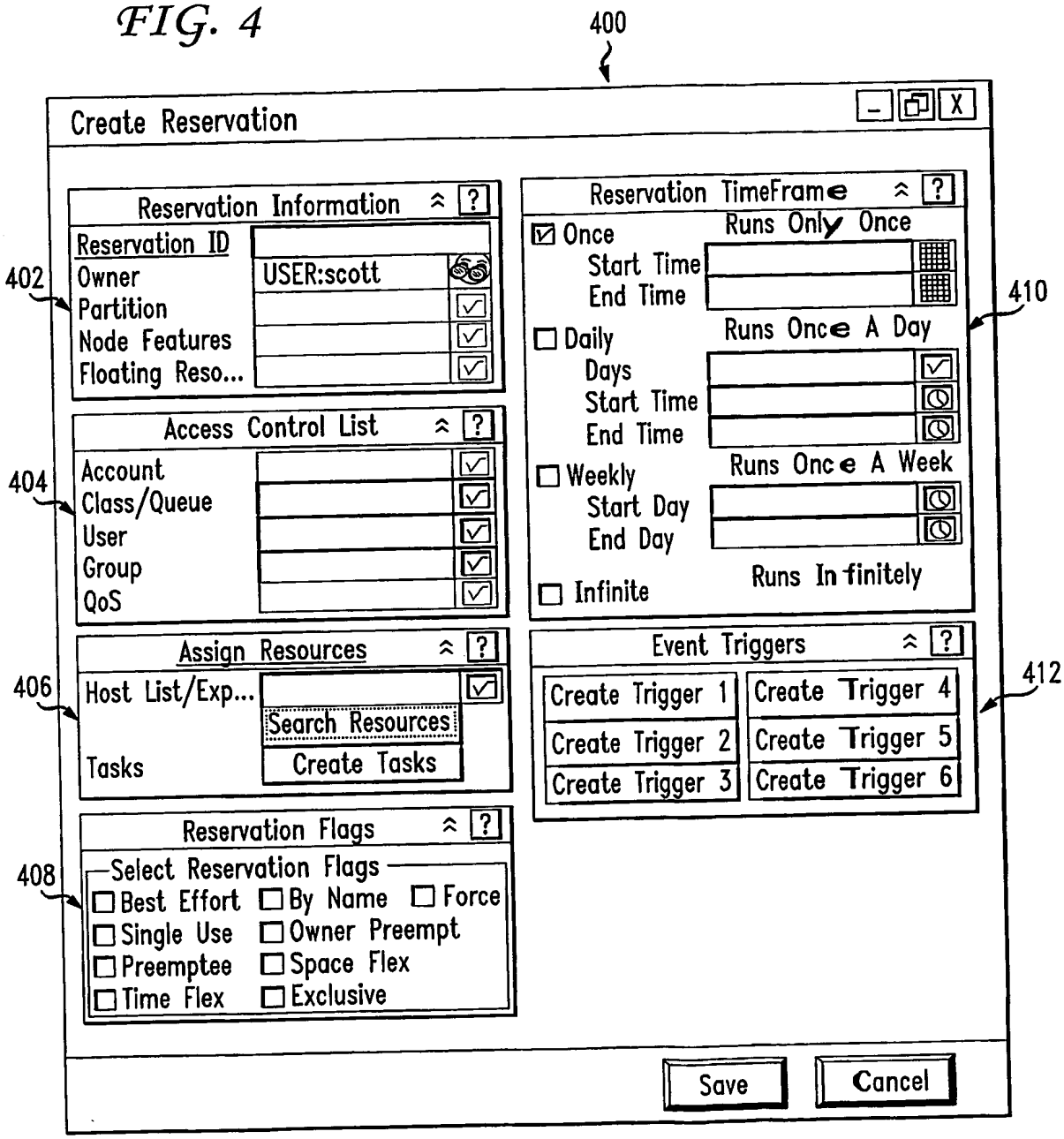

400
Create Reservation
Reservation Information
Reservation ID
Owner
USER:scott
Partition
Node Features
Floating Reso...
402
Access Control List
Account
Class/Queue
User
Group
QoS
404
Assign Resources
Host List/Exp...
Search Resources
Tasks
Create Tasks
406
Reservation Flags
Select Reservation Flags
Best Effort
By Name
Force
Single Use
Owner Preempt
Preemptee
Space Flex
Time Flex
Exclusive
408
Reservation TimeFrame
Once
Runs Only Once
Start Time
End Time
Daily
Runs Once A Day
Days
Start Time
End Time
Weekly
Runs Once A Week
Start Day
End Day
Infinite
Runs Infinitely
410
Event Triggers
Create Trigger 1
Create Trigger 2
Create Trigger 3
Create Trigger 4
Create Trigger 5
Create Trigger 6
412
Save
Cancel

FIG. 5

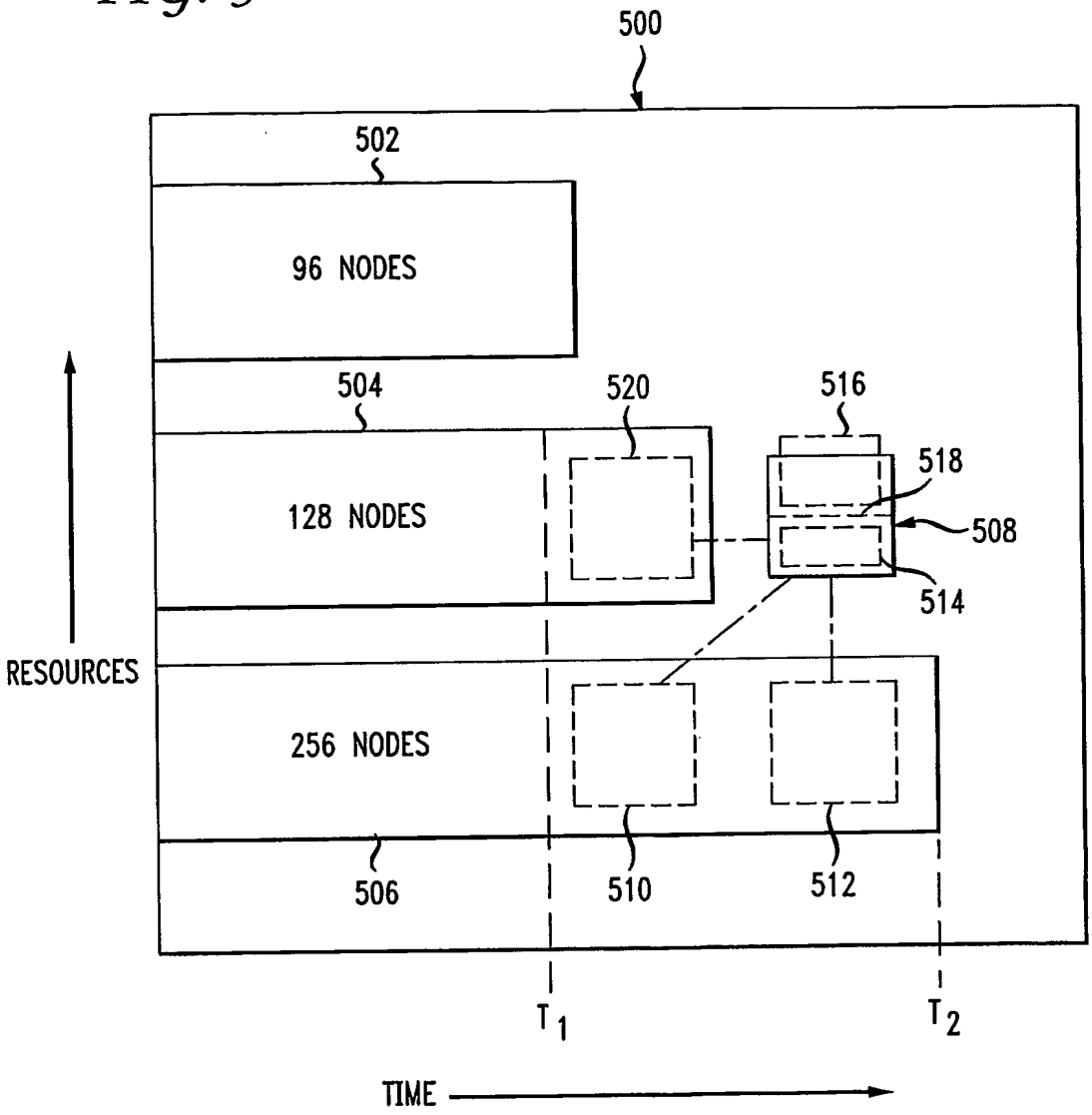

500
502
96 NODES
504
520
516
518
508
128 NODES
514
RESOURCES
256 NODES
506
510
512
T1
T2
TIME

FIG. 6

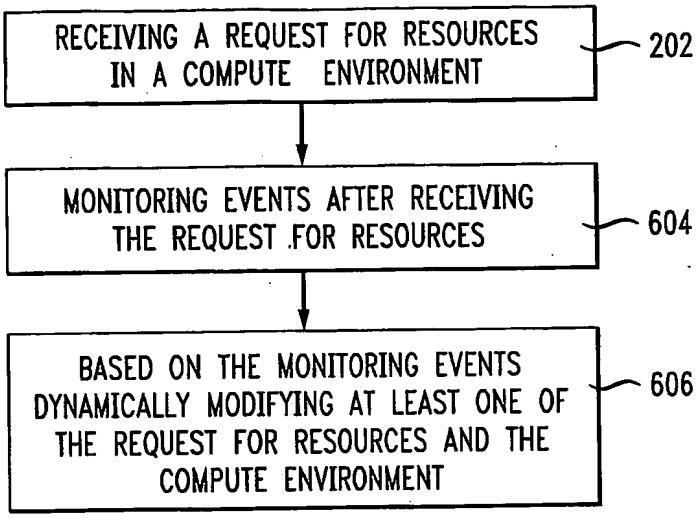
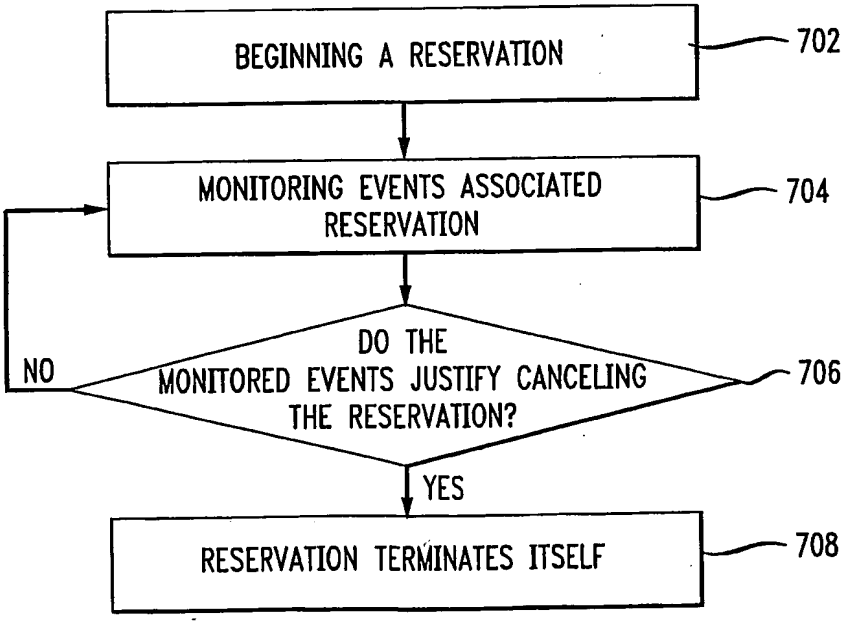

RECEIVING A REQUEST FOR RESOURCES IN A COMPUTE ENVIRONMENT
202
MONITORING EVENTS AFTER RECEIVING THE REQUEST FOR RESOURCES
604
BASED ON THE MONITORING EVENTS DYNAMICALLY MODIFYING AT LEAST ONE OF THE REQUEST FOR RESOURCES AND THE COMPUTE ENVIRONMENT
606

FIG. 7


BEGINNING A RESERVATION
702
MONITORING EVENTS ASSOCIATED RESERVATION
704
DO THE MONITORED EVENTS JUSTIFY CANCELING THE RESERVATION?
706
NO
YES
RESERVATION TERMINATES ITSELF
708